Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Estimating the integrated volatility using high-frequency data with zero durations

Zhi Liu [a], Xin-Bing Kong [b,*], Bing-Yi Jing [c]

[a] University of Macau, Macao
[b] Nanjing Audit University, China
[c] Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 23 October 2016
Received in revised form 16 August 2017
Accepted 3 December 2017
Available online xxxx

JEL classification:
C01
C02
C13
C14
C80

Keywords:
Itô semimartingale
High frequency data
Multiple transactions
Realized power variations
Microstructure noise
Central limit theorem

ABSTRACT

In estimating integrated volatility using high-frequency data, it is well documented that the presence of microstructure noise presents a major challenge. Recent literature has shown that the presence of multiple observations, a common feature in datasets, brings additional difficulty. In this study, we show that the preaveraging estimator is still consistent under multiple observations, and the related asymptotic distribution of the estimator is established. We also show that the preaveraging estimator based on multiple observations achieves the same asymptotic efficiency as the "ideal" estimator that assumes we know the exact trading times of all transactions. Simulation studies support the theoretical results, and we also illustrate the estimator using real data analysis.

1. Introduction

In finance, volatility is a primary component in asset pricing and risk management. High-frequency data have become widely available in recent decades, which have encouraged research on the inference of volatility. As a result, a large number of studies have been devoted to the problem of estimating integrated volatility. One milestone in financial econometrics is the introduction of the concept of realized volatility, which consistently estimates the price variation accumulated over some fixed time interval, such as one day, by summing over the squared high-frequency returns (see Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002a) and Andersen et al. (2005, 2006, 2010), among others).

In practice, empirical studies have shown that the dynamics of (ultra-) high-frequency data largely differ from the semimartingale-type behavior of low-frequency data. The usual understanding of this phenomenon is that the efficient price is contaminated by market microstructure effects, or so-called microstructure noise. Consequently, intra-day asset prices are viewed as noisy observations of the efficient price; hence, the estimation procedure has to be built within this so-called microstructure noise context. Some de-noising methods have been proposed (see, for example, the two-time-scale approach studied by Aït-Sahalia et al. (2005), Mykland and Zhang (2009) and Aït-Sahalia et al. (2010, 2011); the multi-scale method suggested by Zhang (2006); the realized kernel method studied by Barndorff-Nielsen et al. (2008, 2009, 2011); the quasi-maximum likelihood approach proposed by Xiu (2010); the pre-averaging approach studied by Jacod et al. (2009), Li (2013) and Jing et al. (2014); and references therein).

Theoretically, the approaches developed in the above-mentioned literature are only applied to those datasets that contain exactly one transaction during one time stamp, such as, for example, 5-minute, 1-minute, or 5-second returns. This assumption, however, is challenged by a typical stylized characteristic of an ultra-high-frequency dataset. With tick-by-tick transaction data, one time stamp can often include more than one record even if it is only one millisecond long. For example, for the stock of Microsoft Corporation from January 1 to September 16, 2016, the LOBSTER database recorded a total of 6 568 006 transactions in 2 664 323 efficient trading milliseconds, which means that during each of

* Corresponding author.
E-mail addresses: liuzhi@umac.mo (Z. Liu), kongxb@fudan.edu.cn (X.-B. Kong), majing@ust.hk (B.-Y. Jing).

https://doi.org/10.1016/j.jeconom.2017.12.008
0304-4076/

those milliseconds, at least one trade occurs. Among the 2 664 323 efficient trading milliseconds, 1 190 686 ms (about 45%) contain at least two transactions. In an extreme case, 740 transactions were recorded for one millisecond.

Hence, to implement the theoretical approaches, a data-cleaning procedure is necessary, which is also a very important step. Data-cleaning procedures have been addressed by Falkenberry (2001), Hansen and Lunde (2006), Brownlees and Gallo (2006), and Barndorff-Nielsen et al. (2008, 2009). In addition to the other aspects of high-frequency data, when encountering multiple transactions, to the best of our knowledge, researchers have used the following methods to clean the data:

1. use the average of the records within the same time stamp, as, for example, Aït-Sahalia et al. (2010) and Jing et al. (2017);
2. use the median of the prices with an identical time stamp, as, for example, Barndorff-Nielsen et al. (2011);
3. use volume-weighted average prices, as, for example, Christensen et al. (2010);
4. use a single representative price by random drawing or by simply picking the last record, as, for example, Jing et al. (2017).

After the data-cleaning step, the existing approaches could be applied. However, the theoretical accuracy of the estimators based on the specific data-cleaning scheme has not yet been confirmed. Therefore, a natural question arises: if the dataset contains multiple observations, do the estimators of integrated volatility under the above-mentioned data-cleaning procedures remain accurate? To answer this question, we conduct a simple simulation study as follows. With a standard Brownian motion B_t, we let $dX_t = dB_t$ for $t \in [0, 1]$. At the time points $\{t_j = \frac{j}{N}, j = 1, 2, \ldots, N\}$, we generate X_{t_j} and $\epsilon_{t_j} \sim^{iid} N(0, \omega^2)$ with $N = 23,400 \times L$, where L represents the number of multiple observations during a time stamp, ω^2 is the variance of noise, and both will be specified in Fig. 1. Now, we let $\{s_i, i = 0, 1, \ldots, n\}$ with $n = 23,400$ as the recording times, or, in other words, we regard the X_{t_j}'s with $s_{i-1} < t_j \leq s_i$ as the observations at the point s_i. To deal with multiple observations, we apply the first two of the above-mentioned procedures, that is, we use the average of the grouped records and the median of the grouped records. Since the third procedure requires the volume of transactions, we do not consider it. More precisely, the observations at the recording time s_i are $\{X_{t_j}, j = (i-1)L + 1, \ldots, iL\}$, and we denote

$$X^{ave}_{s_i} := \frac{1}{L}\sum_{k=1}^{L} X_{t_{(i-1)L+k}} \quad \text{and}$$

$$X^{med}_{s_i} := \text{median}\{X_{t_j}, j = (i-1)L + 1, \ldots, iL\}.$$

Based on $X^{ave}_{s_i}$ and $X^{med}_{s_i}$, we can employ the existing de-noising methods. We choose the preaveraging approach. Specifically,

$$PA^{ave}_n = \frac{\sqrt{\Delta_n}}{\theta\psi_2}\sum_{i=0}^{n-k_n}(\Delta_{i,k_n}X^{ave})^2 - \frac{\psi_1\Delta_n}{2\theta^2\psi_2}\sum_{i=0}^{n-1}(X^{ave}_{s_{i+1}} - X^{ave}_{s_i})^2,$$

$$PA^{med}_n = \frac{\sqrt{\Delta_n}}{\theta\psi_2}\sum_{i=0}^{n-k_n}(\Delta_{i,k_n}X^{med})^2 - \frac{\psi_1\Delta_n}{2\theta^2\psi_2}\sum_{i=0}^{n-1}(X^{med}_{s_{i+1}} - X^{med}_{s_i})^2,$$

where $\Delta_n = \frac{1}{n}$, $k_n = \lfloor\theta\sqrt{\frac{1}{\Delta_n}}\rfloor$, θ is a constant, $\psi_2 = \int_0^1 g^2(s)ds$, $\psi_1 = \int_0^1 (g'(s))^2 ds$ with $g(x) = \min(x, 1-x)$ and $\Delta_{i,k_n}X^{ave} = \sum_{k=1}^{k_n-1} g(\frac{k}{k_n})(X^{ave}_{s_{i+k+1}} - X^{ave}_{s_{i+k}})$. The $\Delta_{i,k_n}X^{med}$ is similarly defined. To verify the accuracy of the two estimators, we compute the relative bias, and the results are shown in Fig. 1.

From the figure, we can see that the preaveraging estimators based on averages and medians perform well for all cases. That is, the data-cleaning procedures taking either averages or medians do not affect the consistency of the preaveraging estimator. In this study, we theoretically prove the consistency of the first estimator, which is the average over the multiple observations. The asymptotic distribution of the estimator is also obtained. We consider a practically feasible estimator that allows for varying numbers of multiple observations, or, in other words, that the numbers of the multiple observations ($L_i's$) can be different. Moreover, the estimator remains valid for any pattern of latent $t_i's$ within the recording interval $(s^n_{j-1}, s^n_j]$ and allows the presence of jumps in the latent process. The other data-cleaning procedure, using the median of the prices, is also interesting and will be addressed in near future.

The results of this paper extend the existing ones in literature. With the presence of multiple observations, Jing et al. (2017) theoretically studied the estimation of integrated volatility without presence of noise; they suggested a noise-robust estimator of integrated volatility in a simulation example for a special case, that is, $L_i \equiv L$ and $t_j - t_{j-1} \equiv \Delta_n/L$ for $i = 1, \ldots, n$ and $j = 1, \ldots, nL$. Liu (2016) investigated the estimation of co-volatility under multiple observations with the presence of noise, but the study only derived the consistency for the special case of $L_i \equiv L$ as well. Both the results Liu (2016) and Jing et al. (2017) are based on the continuous latent process. Liu (2017) derived the limiting behavior of the multi-power variations, where the result is based on the assumption that the multiple observations equally spaced partition the recording intervals. Therefore, this paper not only extends the setting of the above mentioned studies, but also provides a practical feasible estimator.

The remainder of the paper is organized as follows. Section 2 describes the model setup. In Section 3, we derive the asymptotic results. We compare the asymptotic variance of the proposed estimator to that of other estimators in Section 5. The results of simulation studies are included in Section 6, and applications to real high-frequency data are included in Section 7. We conclude our paper with Section 8, and all technical proofs are in the Appendix.

2. Setup

2.1. The model

Let $\{X_t, t \geq 0\}$ denote the (unobservable) efficient log-price process defined on the probability space $(\Omega^{(0)}, \mathcal{F}^{(0)}, P^{(0)})$ equipped with filtration $\{\mathcal{F}^{(0)}_t\}_{t\geq 0}$. It is well-known that, under the no-arbitrage assumption, security price processes must follow a semimartingale (see, e.g., Delbaen and Schachermayer (1994)). We assume that X follows a one-dimensional semimartingale of the form

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \tag{1}$$

where W is a standard Brownian motion, $\{b_s, s \geq 0\}$ is a locally bounded adapted process, and $\{\sigma_s, s \geq 0\}$ is an adapted càdlàg process.

The process given in (1) is a rather canonical model in finance due to the fact that all continuous local martingales with absolutely continuous quadratic variation can be written in the form of (1). Since $\{\sigma_s, s \geq 0\}$ is càdlàg, all powers of σ are locally integrable with respect to the Lebesgue measure. In particular, we have $C_t := \int_0^t \sigma_s^2 ds < \infty$. Moreover, both $\{b_s, s \geq 0\}$ and $\{\sigma_s, s \geq 0\}$ can have, for example, jumps, intra-day seasonality, and long memory.

Throughout, $\kappa(t, x)$ will denote a continuous truncation function, that is, a continuous function with bounded support such that

Please cite this article in press as: Liu Z., et al., Estimating the integrated volatility using high-frequency data with zero durations. Journal of Econometrics (2018), https://doi.org/10.1016/j.jeconom.2017.12.008.